KnowEng, a Scalable Knowledge Engine for Large-Scale Genomic Data

The University of Illinois Urbana-Champaign

PIs: Jiawei Han, Saurabh Sinha, Jun Sorg, and Richard Weinshilboum Grant Number: 1-U54GM114838-01

The primary goal of the proposed Center of Excellence is to build a powerful and scalable Knowledge Engine for Genomics, KnowEnG. KnowEnG will transform the way biomedical researchers analyze their genome-wide data by integrating multiple analytical methods derived from the most advanced data mining and machine learning research to use the full breadth of existing knowledge about the relationships between genes as background, and providing an intuitive and professionally designed user interface. In order to achieve these goals, the project includes the following components: (1) gathering and integrating existing knowledgebases documenting connections between genes and their functions into a single Knowledge Network; (2) developing computational methods for analyzing genome-wide user datasets in the context of this pre-existing knowledge; (3) implementing these methods into scalable software components that can be deployed in a public or private cloud; (4) designing and implementing a Web-based user interface, based on the HUBZero toolkit, that enables the interactive analysis of user-supplied datasets in a graphics-driven and intuitive fashion; (5) thoroughly testing the functionality and usefulness of the KnowEnG environment in three large scale projects in the clinical sciences (pharmacogenomics of breast cancer), behavioral sciences (identification of gene regulatory modules underlying behavioral patterns) and drug discovery (genome-based prediction of the capacity of microorganisms to synthesize novel biologically active compounds). The KnowEng environment will be deployed in a cloud infrastructure and fully available to the community, as will be the software developed by the Center. The proposed Center is a collaboration between the University of Illinois (UIUC), a recognized world leader in computational science and engineering, and the Mayo Clinic, one of the leading clinical care and research organizations in the worid, and will be based at the UIUC Institute for Genomic Biology, which has state-of-the-art facilities and a nationally recognized program of multidisciplinary team-based genomic research. PUBLIC HEALTH RELEVANCE: Physicians and biologists are now routinely producing very large, genome-wide datasets. These data need to be analyzed in the context of an even larger corpus of publically available data, in a manner that is approachable to non-specialist doctors and scientists. The proposed Center will leverage the latest computational techniques used to mine corporate or Internet data to enable the intuitive analysis and exploration of biomedical Big Data.